Project 1: Exploratory Data Analysis, Wrangling, and Feature Engineering on the ADNI Dataset

Josiah Chuku

Institution: Florida A&M University

Department of Computer Science

February 10, 2026

1 Project Overview

This technical report details the data preparation and engineering pipeline applied to the ADNI (`TADPOLE_D3`) clinical dataset. The objective is to transform 383 high-dimensional features into a clean, scaled, and reduced format suitable for machine learning, while justifying every methodological decision as per the requirements of the course.

2 Task 1: Data Cleaning

2.1 1.1 Missing Data Analysis

The dataset consists of 896 observations. An initial scan revealed significant missing data across several neuroimaging and clinical markers. Figure 1 visualizes the sparsity of the data.

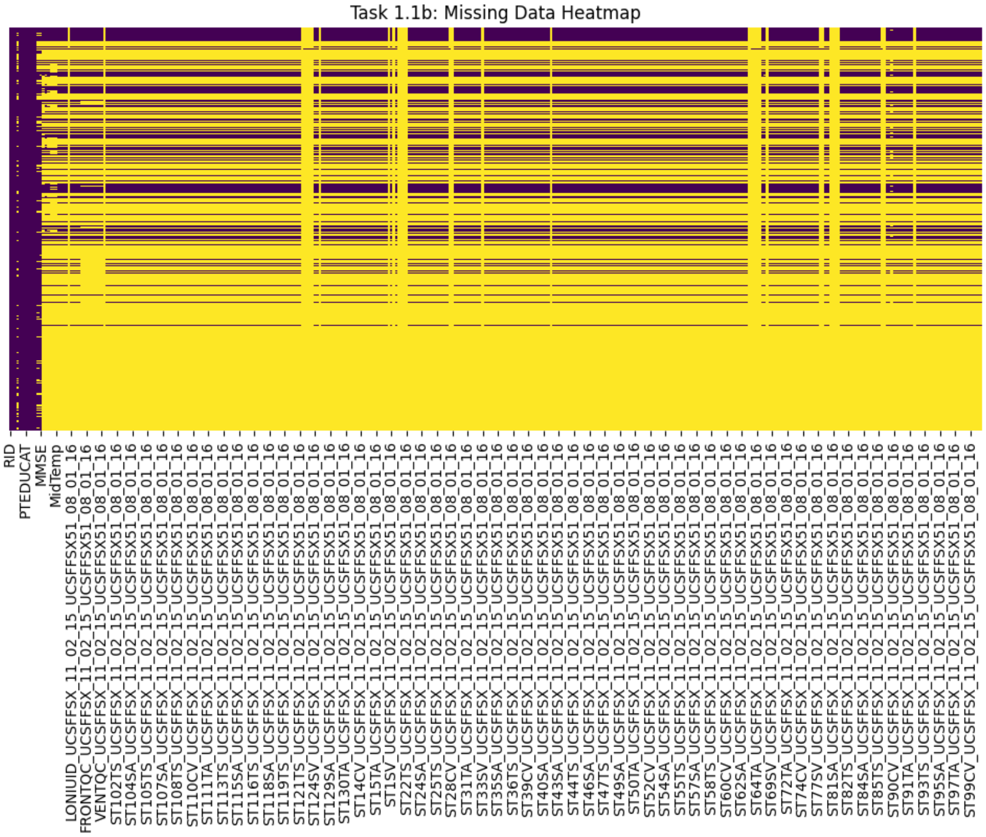

Figure 1: Heatmap of missing data across 383 features (Yellow indicates missing values).

Justification for Decision:

- **Dropping Columns:** Features with $> 50\%$ missing data were removed. Retaining these would require extreme imputation, which risks introducing significant bias and "hallucinated" patterns.

- **Median Imputation:** For the remaining numerical features, **Median Imputation** was chosen over Mean Imputation. Clinical data (like brain volumes) often contain outliers; the median is more robust and prevents extreme values from skewing the representative center of the distribution.

2.2 1.2 Incorrect Data Handling

Impossible values were detected in the neuroimaging columns (prefix `ST`).

- **Detection:** Logic was implemented to find non-positive values (≤ 0) in brain volume columns, which are biologically impossible.

- **Action:** These values were treated as "garbage data," converted to NaNs, and handled via the imputation pipeline to ensure algorithmic stability.

2.3 1.3 Unnecessary Data

Features such as RID, PTID, SITE, and EXAMDATE were identified as unnecessary. While useful for database management, these are unique identifiers that do not represent generalizable biological patterns. Including them would lead the model to ”memorize” specific patients rather than learn predictive features.

3 Task 2: Feature Engineering

3.1 2.1 Encoding Categorical Features

The categorical features (e.g., DX for Diagnosis) were processed using **One-Hot Encoding**. **Justification:** Diagnosis is a nominal category. Using Ordinal Encoding (1, 2, 3) would incorrectly imply that one diagnosis is mathematically ”greater” than another. One-Hot Encoding creates binary columns, allowing the model to treat each state independently, though it slightly increases dimensionality.

3.2 2.2 Scaling Comparison and Selection

Scaling was performed on a sample of 50 observations for three features: AGE, PTEDUCAT, and ADAS13.

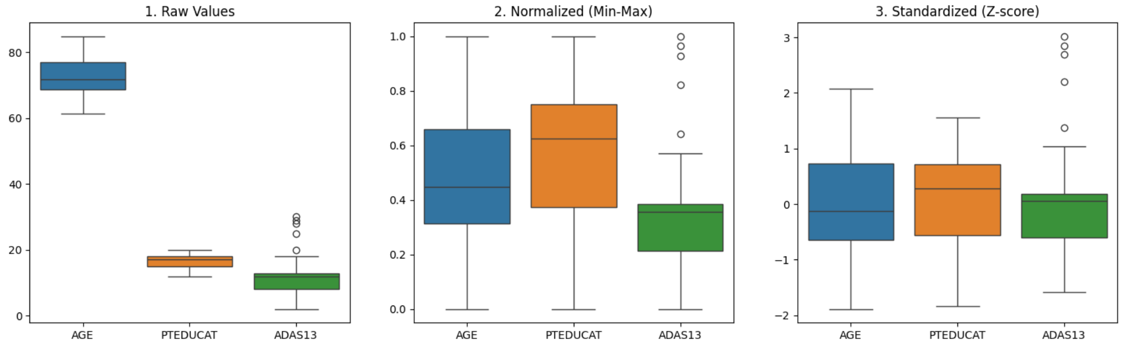

Figure 2: Comparison of Raw, Normalized (Min-Max), and Standardized (Z-score) values.

Justification: Standardization (Z-score) was selected as the final method. Unlike Normalization, which squashes data into a $0 - 1$ range, Standardization centers the data at a mean of 0 with unit variance. This is essential for the Principal Component Analysis performed in the next step.

4 Task 3: Dimensionality Reduction

4.1 3.1 The Curse of Dimensionality

The ADNI dataset suffers from the *Curse of Dimensionality*. With 383 features, the ”volume” of the feature space is vast, making the 896 data points extremely sparse. This sparsity increases the risk of overfitting, as models can easily find random noise that appears to be a trend.

4.2 3.2 Principal Component Analysis (PCA)

PCA was implemented to project the high-dimensional space into two components.

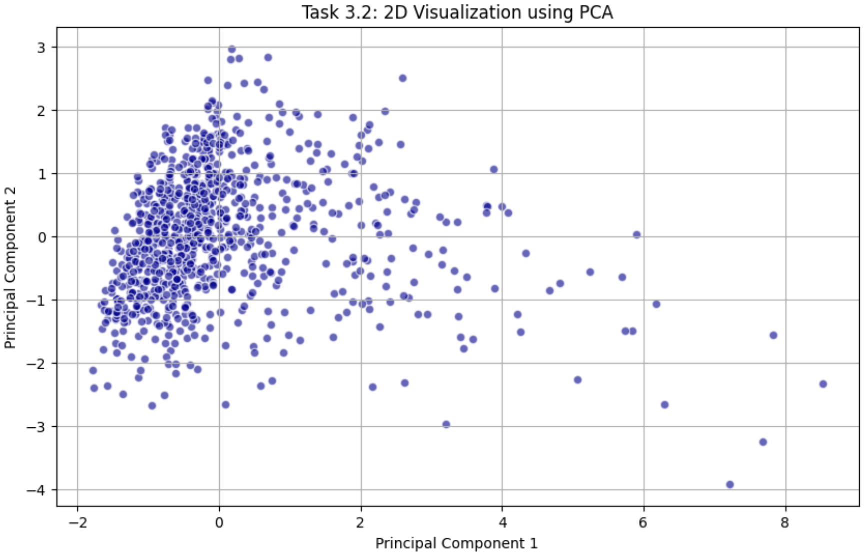

Figure 3: 2D Visualization of the ADNI dataset after PCA dimensionality reduction.

Conceptual Explanation: PCA works by rotating the coordinate axes to align with the directions of maximum variance (Principal Components). By keeping only the first two components, we retain the majority of the "information" while discarding redundant or highly correlated dimensions, effectively mitigating the curse of dimensionality.